Final

Aritzi Piedras Silva

2/26/2020

# Accidental Drug Related Deaths 2012-2018

This data was provided by the state of Connecticut’s Office of the Chief Medical Examiner.

##### The following packages will be used in this analysis:

Tidyverse

Caret

Rpart

Rpart.plot

## Connecticut Overdose Statistical Analysis

This data was provided by the state of Connecticut’s Office of the Chief Medical Examiner.

### Introduction

Accidental death by drug overdose has clarinet thousands of lives, what can we do to prevent more?

Each year Opioids, prescription pain medications and many more other drugs claim countless lives due to accidental overdose. Unfortunately, the overdose epidemic has become a public health crisis nation-wide. This project will focus on shining light on the many deaths that drug abuse has claimed in the state of Connecticut from the years 2012 to 2018. I will be analyzing data collected by The Office of the Chief Medical Examiner of accidental overdose deaths in order to come up with better solution as to how we as a society can come up with better ways to prevent more lives being claimed by overdose. This data contains different variables that could help us predict the average number of drugs found in the body at the time of death depending on race, gender, and age.

**Research questions:**

How heavily does age affect the likelihood of someone that will die from accidental overdose?

Do men or women die more from accidental overdose?

What is the average number of drugs found in a body at the time of death depending on age, sex, and race?

Does race play a factor in drug addiction?

**Plan of problem statement:** In this research analysis, I will be exploring the variables of age, race, gender, and drug categories to best predict the average number of drugs a human can intake that causes overdose. Additionally, I will be plotting on different bar-graphs of gender, age, and race to compare which groups are most affected by this epidemic.

This approach will partially cover this problem. Unfortunately, the number of overdose deaths involving opioids, prescription, and non-prescription, continue to increase yearly. Thus, this dataset only included data from the state of Connecticut rather than nationwide, making it hard to predict the number of drugs used that has caused humans to lose their battle against addiction nationwide or globally. Additionally, this dataset does not contain the socio-economic status of each individual, therefore predicting which social classes are most affected will not be able to be done. However, this analysis could potentially shine a light on general groups suffering addiction in Connecticut based on age, gender, and race to regulate and prevent more deaths.

drugdf<- read\_csv("https://data.ct.gov/api/views/rybz-nyjw/rows.csv?accessType=DOWNLOAD", quoted\_na = TRUE, trim\_ws = TRUE)

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## Age = col\_double()  
## )

## See spec(...) for full column specifications.

spec(drugdf)

## cols(  
## ID = col\_character(),  
## Date = col\_character(),  
## DateType = col\_character(),  
## Age = col\_double(),  
## Sex = col\_character(),  
## Race = col\_character(),  
## ResidenceCity = col\_character(),  
## ResidenceCounty = col\_character(),  
## ResidenceState = col\_character(),  
## DeathCity = col\_character(),  
## DeathCounty = col\_character(),  
## Location = col\_character(),  
## LocationifOther = col\_character(),  
## DescriptionofInjury = col\_character(),  
## InjuryPlace = col\_character(),  
## InjuryCity = col\_character(),  
## InjuryCounty = col\_character(),  
## InjuryState = col\_character(),  
## COD = col\_character(),  
## OtherSignifican = col\_character(),  
## Heroin = col\_character(),  
## Cocaine = col\_character(),  
## Fentanyl = col\_character(),  
## FentanylAnalogue = col\_character(),  
## Oxycodone = col\_character(),  
## Oxymorphone = col\_character(),  
## Ethanol = col\_character(),  
## Hydrocodone = col\_character(),  
## Benzodiazepine = col\_character(),  
## Methadone = col\_character(),  
## Amphet = col\_character(),  
## Tramad = col\_character(),  
## Morphine\_NotHeroin = col\_character(),  
## Hydromorphone = col\_character(),  
## Other = col\_character(),  
## OpiateNOS = col\_character(),  
## AnyOpioid = col\_character(),  
## MannerofDeath = col\_character(),  
## DeathCityGeo = col\_character(),  
## ResidenceCityGeo = col\_character(),  
## InjuryCityGeo = col\_character()  
## )

## Race Analysis

#Race Data set   
summarydf<- drugdf %>%   
 mutate(other = case\_when(  
 Race == "Asian, Other" ~ "Other",  
 Race == "Asian Indian" ~ "Other",  
 Race == "Chinese" ~ "Other",  
 Race == "Native American, Other" ~ "Other",  
 Race == "Unknown" ~ "Other",  
 Race == "Other" ~ "Other",  
 Race == "NA" ~ "Other",  
 Race == "Black" ~ "Black",  
 Race == "Hispanic, White" ~ "Hispanic",  
 Race == "Hispanic, Black" ~ "Hispanic",  
 Race == "White" ~ "White")) %>%  
 pivot\_longer(cols = OtherSignifican:AnyOpioid, names\_to = "drug", values\_to = "Taken") %>%   
 group\_by(other, drug, Taken)%>%  
 filter(drug %in% c("Heroin", "Cocaine", "Fentanyl", "Oxycode", "Ethanol", "Hydrocodone", "Oxycodone", "Oxymorphone", "Ethanol",  
 "Benzodiazepine", "Methadone","Tramad", "Amphet"), !is.na(Taken), !is.na(other)) %>%  
 summarize(counts.drug = n())  
  
glimpse(summarydf)

## Observations: 48  
## Variables: 4  
## Groups: other, drug [44]  
## $ other <chr> "Black", "Black", "Black", "Black", "Black", "Black", "Bl…  
## $ drug <chr> "Amphet", "Benzodiazepine", "Cocaine", "Ethanol", "Fentan…  
## $ Taken <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y…  
## $ counts.drug <int> 8, 48, 235, 126, 182, 165, 5, 24, 35, 1, 11, 9, 94, 210, …

#### Summarydf:

In **summarydf**, I separated race into different categories –

Asian, other, Asian Indian, Chinese, Native American, Unknown and Other were placed into “Other” category.

Hispanic – White and Hispanic Black were placed under “Hispanic”

Black was placed under “Black”

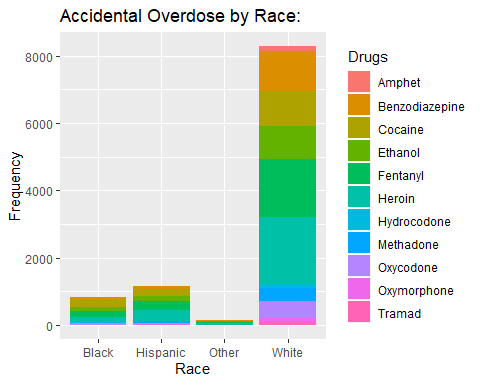
White was placed under “White”

Additionally, I grouped the following different drugs into one category “drug”: Heroin, Cocaine, Fentanyl, Oxycodone, Ethanol, Hydrocodone, Oxycodone, Oxymorphone, Ethanol, Benzodiazepine, Methadone, Tramadol, Amphetamine. I removed the NAs from the created attributes by using: “!is.na(x)”.

## 

## Bar Graph of Accidental Overdose by Race

#Race bargraph  
ggplot(summarydf, aes(x = factor(other), y = counts.drug, fill = factor(drug)))+geom\_bar(stat = "identity", position = "stack") + labs(title = "Accidental Overdose by Race: ", x = "Race", y= "Frequency") +   
scale\_fill\_discrete(name = "Drugs")



This Bar-graph demonstrates the significance of the white race and drug use. Out of the four groups, the White race has had the most deaths due to drug abuse. Heroin has claimed most deaths amongst all four groups followed by Fentanyl despite age or gender. I’m extremely intrigued to find out why the white race has had more causalities with drug addiction compared to the other race categories.

## Age Analysis

#Age data   
AgevsdrugAnalysis<- drugdf %>%   
 mutate(Agegroup = case\_when(  
 Age %in% c(0:17) ~ "<17",  
 Age %in% c(18:30) ~ "18-30",  
 Age %in% c(31:49) ~ "31-49",  
 Age %in% c(50:68) ~ "50-68",  
 Age > 69 ~ ">69",   
 TRUE~ "Unknown"), Agegroup = factor(Agegroup, labels = c("<17","18-30", "31-49", "50-68", ">69", "Unknown"))) %>%  
 group\_by(Agegroup) %>%  
 pivot\_longer(cols = OtherSignifican:AnyOpioid, names\_to = "drug", values\_to = "Taken") %>%   
 group\_by(drug, Agegroup) %>%  
 filter(drug %in% c("Heroin", "Cocaine", "Fentanyl", "Oxycode", "Ethanol", "Hydrocodone", "Oxycodone", "Oxymorphone", "Ethanol",  
 "Benzodiazepine", "Methadone","Tramad", "Amphet"), !is.na(Taken), !is.na(Agegroup)) %>%  
 summarize(counts.drug = n())  
  
glimpse(AgevsdrugAnalysis)

## Observations: 66  
## Variables: 3  
## Groups: drug [11]  
## $ drug <chr> "Amphet", "Amphet", "Amphet", "Amphet", "Amphet", "Amphet…  
## $ Agegroup <fct> <17, 18-30, 31-49, 50-68, >69, Unknown, <17, 18-30, 31-49…  
## $ counts.drug <int> 1, 0, 41, 81, 36, 0, 4, 10, 245, 611, 468, 5, 2, 8, 271, …

In **AgevsdrugAnalysis**, I separated ages into different groups:

Ages 0-17 was grouped to ”<17”

Ages 18 -30 was grouped to “18-30”

Ages 31-49 was grouped to “31-49”

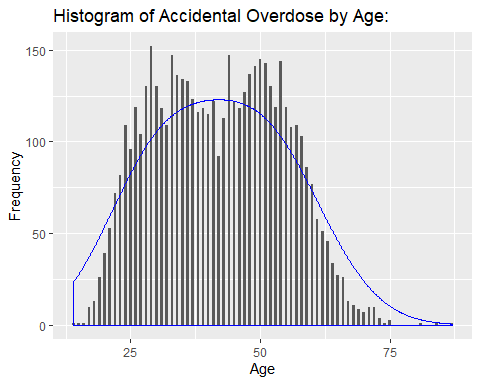
Ages 50 to 68 was grouped to “50-68”

Ages 69 and up was grouped to “>69”

As well as the list of drugs used by the age groups similar to summarydf

### Histogram of Age

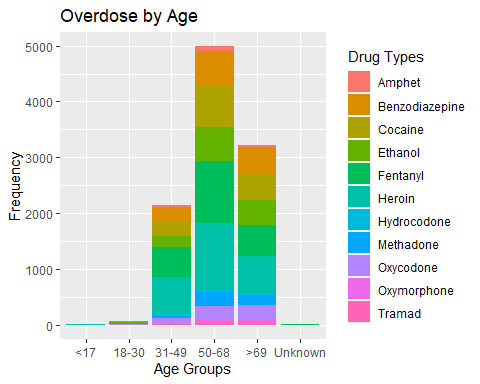
#histogram  
ggplot(drugdf, aes(Age)) + geom\_histogram(binwidth= .5, na.rm = TRUE) +  
 labs(title = "Histogram of Accidental Overdose by Age: ", x = "Age", y = "Frequency") + geom\_density(aes(y=..count..), colour="blue", adjust = 4, na.rm = TRUE)



In this histogram of Age, I can argue that the data of Age is, in fact, double-peaked or bimodal. There are multiple peaks around the late 20s and early 40s. This graph resembles almost two symmetric/bell-shaped histograms as if they have meshed together.

### Bar-plot of Accidental Deaths by Age groups

#age plot  
ggplot(AgevsdrugAnalysis, aes(Agegroup, counts.drug, fill = factor(drug))) + geom\_bar(stat = "identity", position = "stack") +   
 labs(title = "Overdose by Age", x = "Age Groups", y= "Frequency") +   
 scale\_fill\_discrete(name = "Drug Types")



After looking at the histogram of ages, I decided to look at these ages in groups. I noticed the group under “Unknown” which triggered red flags. After looking at the original data, I noticed that “other” was anyone unable to be identified. Therefore, it is classified under “unknown.”

In the age group of 50 to 68, seemed to have a higher overdose rate compared to any other age group despite gender or race. However, every group had a similar pattern being heroin and fentanyl as the drugs with higher usage.

## Gender Analysis

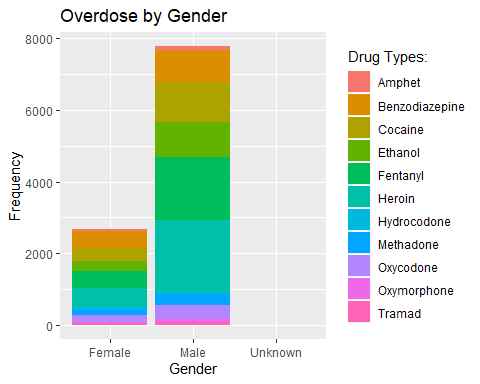
#Gender data   
GenderAnalysis<- drugdf %>%   
 pivot\_longer(cols = OtherSignifican:AnyOpioid, names\_to = "drug", values\_to = "Taken") %>%   
 filter(drug %in% c("Heroin", "Cocaine", "Fentanyl", "Oxycode", "Ethanol", "Hydrocodone", "Oxycodone", "Oxymorphone", "Ethanol",  
 "Benzodiazepine", "Methadone","Tramad", "Amphet"), !is.na(Taken), !is.na(Sex)) %>%  
 group\_by(Sex, Taken, drug) %>%  
 summarize(counts.drug = n())  
  
glimpse(GenderAnalysis)

## Observations: 29  
## Variables: 4  
## Groups: Sex, Taken [7]  
## $ Sex <chr> "Female", "Female", "Female", "Female", "Female", "Female…  
## $ Taken <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y…  
## $ drug <chr> "Amphet", "Benzodiazepine", "Cocaine", "Ethanol", "Fentan…  
## $ counts.drug <int> 48, 463, 381, 274, 465, 539, 55, 159, 204, 35, 46, 1, 1, …

In **GenderAnalysis**, I created this set similar to the previous sets. I compiled the drugs I wanted to demonstrate and how it impacted gender while omitting any NA’s.

## Bar Plot of Gender

#barplot of gender  
ggplot(GenderAnalysis, aes(Sex, counts.drug, fill = factor(drug))) + geom\_bar(stat = "identity", position = "stack") +  
 labs(title = "Overdose by Gender", x = "Gender", y= "Frequency") +   
 scale\_fill\_discrete(name = "Drug Types:")



This bar graph demonstrates that Males had higher drug usage and overdose compared to Females. Similar to the previous plots, Heroin and Fentanyl were most abused by both genders. The unknown is represented by those who were not able to be identified.

### Average number of drug found in a body at the time of death depending on race, gender and age.

overdose <- read.csv("C:/Users/Aritzi Silva/Desktop/drugs.csv")  
# Gender  
overdose$Sex <- ifelse(overdose$Sex == "Male",1,0)  
  
  
# Race   
overdose <- overdose %>%   
 mutate(Race = case\_when(  
 Race == "Asian, Other" ~ "Other",  
 Race == "Asian Indian" ~ "Other",  
 Race == "Chinese" ~ "Other",  
 Race == "Native American, Other" ~ "Other",  
 Race == "Unknown" ~ "Other",  
 Race == "Other" ~ "Other",  
 Race == "NA" ~ "Other",  
 Race == "Black" ~ "African American",  
 Race == "Hispanic, White" ~ "Hispanic",  
 Race == "Hispanic, Black" ~ "Hispanic",  
 Race == "White" ~ "White"))  
  
#Creating Variables for Drugs   
overdose$Heroin <- ifelse(overdose$Heroin == "Y",1,0)  
overdose$Cocaine <- ifelse(overdose$Cocaine == "Y",1,0)  
overdose$Fentanyl <- ifelse(overdose$Fentanyl == "Y",1,0)  
overdose$FentanylAnalogue <- ifelse(overdose$FentanylAnalogue == "Y",1,0)  
overdose$Oxycodone <- ifelse(overdose$Oxycodone == "Y",1,0)  
overdose$Oxymorphone <- ifelse(overdose$Oxymorphone == "Y",1,0)  
overdose$Ethanol <- ifelse(overdose$Ethanol == "Y",1,0)  
overdose$Hydrocodone <- ifelse(overdose$Hydrocodone == "Y",1,0)  
overdose$Benzodiazepine <- ifelse(overdose$Benzodiazepine == "Y",1,0)  
overdose$Methadone <- ifelse(overdose$Methadone == "Y", 1,0)  
overdose$Amphet<- ifelse(overdose$Amphet == "Y",1,0)  
overdose$Tramad <- ifelse(overdose$Tramad == "Y",1,0)  
overdose$Morphine\_NotHeroin <- ifelse(overdose$Morphine\_NotHeroin == "Y",1,0)  
overdose$Hydromorphone <- ifelse(overdose$Hydromorphone == "Y",1,0)  
overdose$Other <- ifelse(overdose$OpiateNOS == "Y", 1,0)  
   
#Drug Total   
  
  
overdose$drugtot <- overdose$Heroin + overdose$Fentanyl + overdose$FentanylAnalogue + overdose$Oxycodone + overdose$Oxymorphone + overdose$Ethanol + overdose$Hydrocodone + overdose$Benzodiazepine + overdose$Methadone + overdose$Amphet + overdose$Tramad +   
 overdose$Morphine\_NotHeroin + overdose$Hydromorphone + overdose$Other

## Total Number of Drugs

summary(overdose$drugtot)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 1.858 2.000 6.000

In this summary, im able to see the median, mean and max drugs used for overdose.

### RMSE

I will be comparing RSEM from a Generalized Linear Model, Random Forest algorithm and a Decision tree in order to best predict the number of drugs that will be found in a system according to age, race,and sex.

# removing missing values   
missing.age <- which(is.na(overdose$Age))  
missing.sex <- which(is.na(overdose$Sex))  
missing.Race <- which(is.na(overdose$Race))  
# train/test datasets   
overdose <- overdose[-c(missing.age, missing.sex,missing.Race),]  
index <- createDataPartition(overdose$drugtot, p= .7, list=FALSE)  
train <- overdose[index,]  
test <- overdose[-index,]

## Generalized Linear Model

#GLM Prediction  
lmod <- glm(drugtot ~ Age + Sex + Race, family = "poisson" , data= train)  
lmod.predict <- predict(lmod,test)  
RMSE(test$drugtot,lmod.predict)

## [1] 1.597541

The Generalized Linear Model demonstrates a RMSE of 1.597541

##### Summary of GLM

summary(lmod)

##   
## Call:  
## glm(formula = drugtot ~ Age + Sex + Race, family = "poisson",   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.99583 -0.71429 0.05266 0.38976 2.56791   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.354343 0.069910 5.069 4.01e-07 \*\*\*  
## Age 0.001148 0.001007 1.140 0.2543   
## Sex 0.011607 0.028252 0.411 0.6812   
## RaceHispanic 0.135833 0.059329 2.289 0.0221 \*   
## RaceOther 0.222760 0.109256 2.039 0.0415 \*   
## RaceWhite 0.239217 0.048213 4.962 6.99e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 2306.2 on 3563 degrees of freedom  
## Residual deviance: 2275.3 on 3558 degrees of freedom  
## AIC: 10680  
##   
## Number of Fisher Scoring iterations: 5

After running the summary on my GLM, i noticed that the Race[Black] wasn’t on the coefficient list. Therefore, black is the baseline and all these coefficient are relative to the baseline. Additionally, Sex has a P value of .68 which is high compared to the other coeffiecients.

## Random Forest Model

#random forest model   
rf <- train(drugtot ~ Age + Sex + Race, method ="rf", data = train)  
rf.predict <- predict(rf,test)  
RMSE(test$drugtot,rf.predict)

## [1] 1.004472

The Random Forest Model demonstrates a RMSE of 1.004472, making it the best model for prediction since it has the lowest Root Means Square Error.

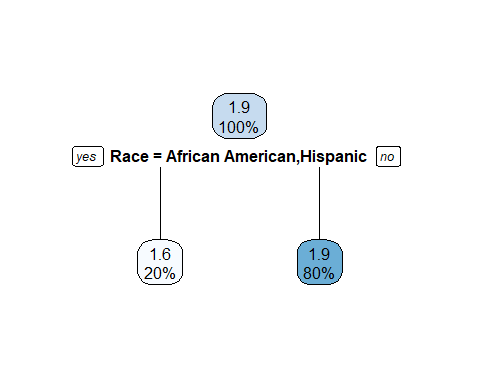
##Decision Tree

#Decision tree  
tree.fit <- rpart(drugtot ~ Age + Sex + Race, data = train)  
tree.predict <- predict(tree.fit,test)  
RMSE(test$drugtot,tree.predict)

## [1] 1.010074

The Decision Tree demonstrates a RMSE of 1.010074

#plot of decision tree  
rpart.plot(tree.fit)



This decision tree deems African American and Hispanic to be the only variables that are most important to split. This tree argues that if you are African American or Hispanic, 20% took 1.5 drugs while 80% took 2 drugs.